



# Introduction to microarray technology and data analysis

**Aron Eklund**

eklund@cbs.dtu.dk

Cancer Systems Biology group  
Center for Biological Sequence Analysis  
DTU Systems Biology

Introduction to Systems Biology  
February 16, 2015

# What is the role of the microarray in **systems biology**?

- The gene expression microarray was the first tool to **efficiently** and **quantitatively** characterize the **global** state of a biological system
- Being replaced by sequencing (RNA-seq)... but many of the same principles apply

# Learning objectives

1. Describe a *gene expression microarray* and what it measures
2. Explain the steps needed to generate gene expression microarray data
3. Explain the steps needed to generate gene expression microarray data
4. Describe at least one application of a gene expression microarray
5. Use R at a basic level
6. Explain and calculate a  $\log_2$  ratio
7. Explain and calculate a  $P$  value using the  $t$  test

# “Expression” of a gene



**Gene expression** = how much  
of a certain *mRNA* is present

**Protein expression** = how much  
of a certain *protein* is present

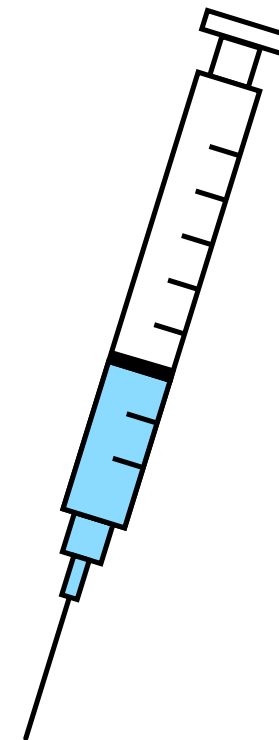
# Why do we want to measure gene expression?

## 1. Biological / medical research

- inference of the function of a gene
- understanding gene regulation
- biomarker discovery

## 2. Clinical diagnostics

- predictive biomarkers for cancer
- identifying cancers of unknown primary



# Why do we want to measure *gene* expression (vs. *protein* expression)?

**Because gene expression is cheaper / faster / easier**

- **Gene expression microarrays are based on sequence-specific hybridization to complementary sequences**

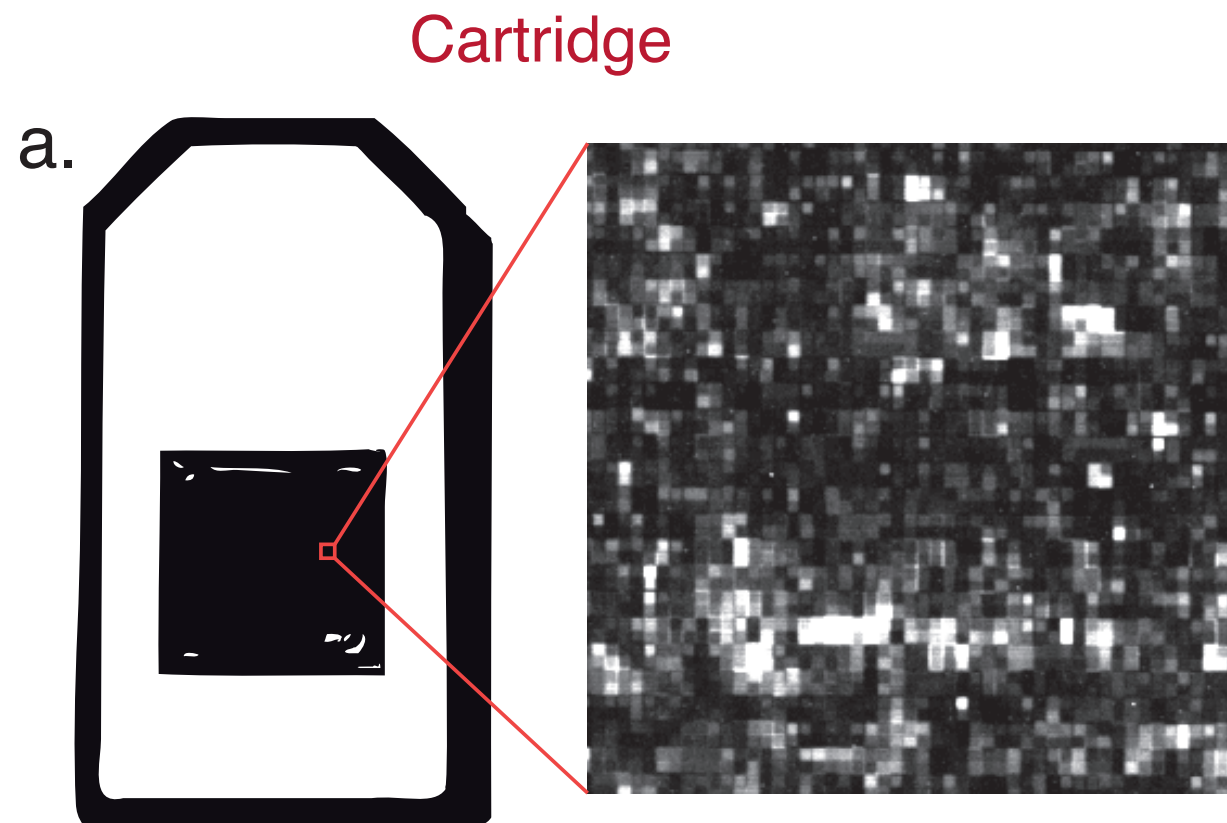
*There is no such thing as a “complementary amino acid sequence”*

- **Gene expression microarrays can be fabricated in parallel**

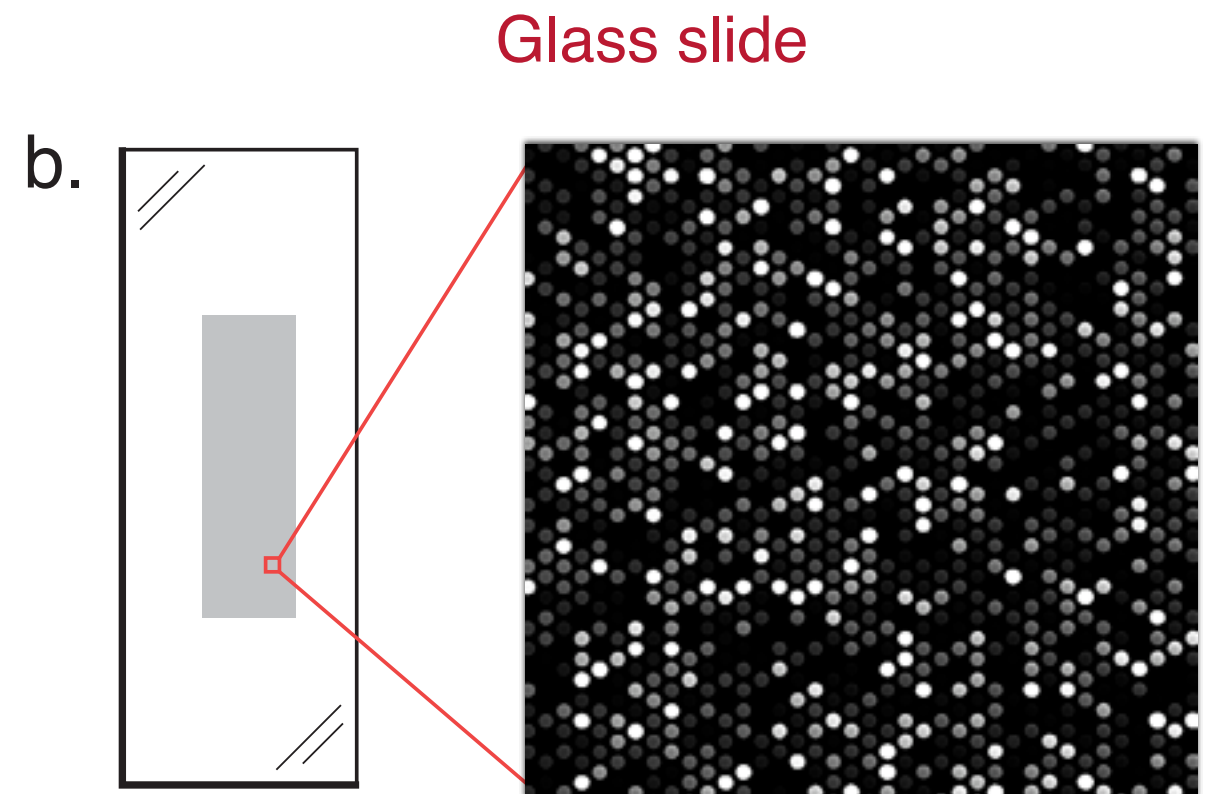
*Protein expression arrays depend on antibodies that must be individually created.*

# The DNA microarray (MICROscopic Array)

Thousands to millions of DNA “spots” in a few square centimeters



In situ photolithographic  
synthesis

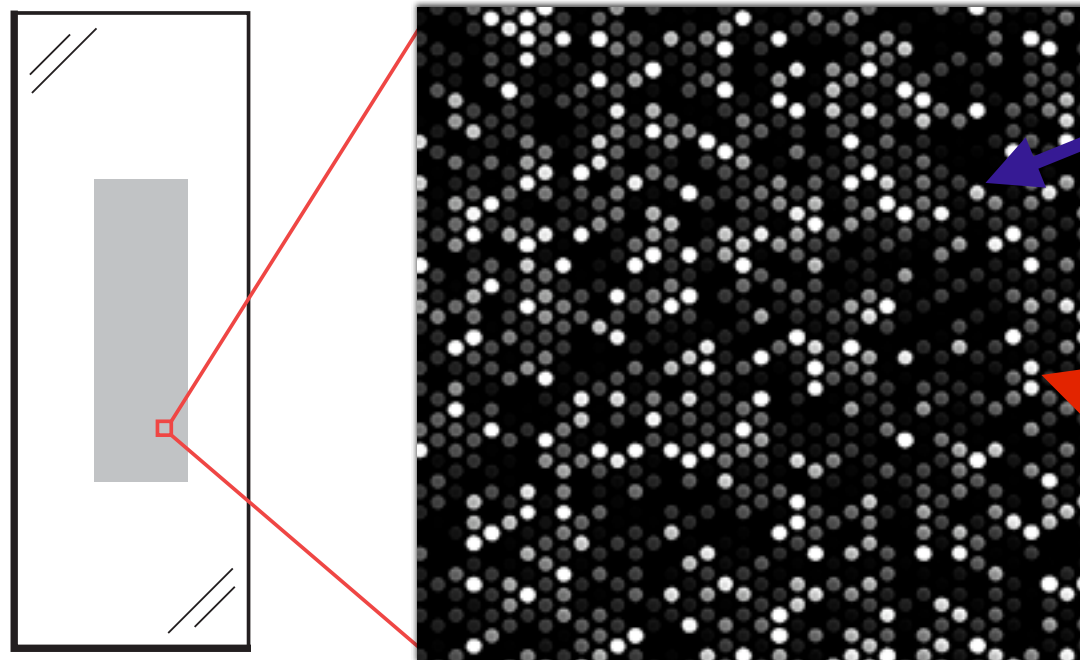


Robotic spotting or  
in situ synthesis

# The DNA microarray

Millions of single-stranded DNA “probes”,  
each with the same sequence:

TCAGCTAGCTAGTCGATGCTAGTCGACGGG



Millions of single-stranded DNA  
“probes” with the same sequence:  
CCGATGCTAGTTCAGCTAGCTAGCGACGATA

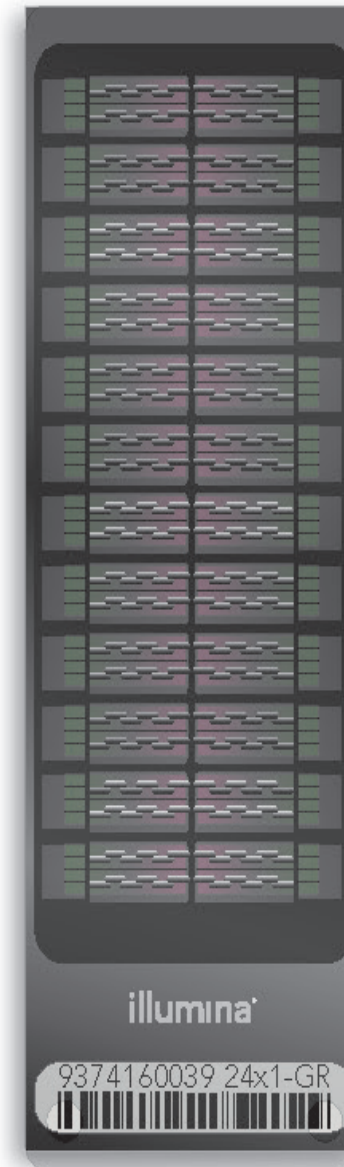
Each “spot” or “feature” is  
designed to detect



# Example microarrays



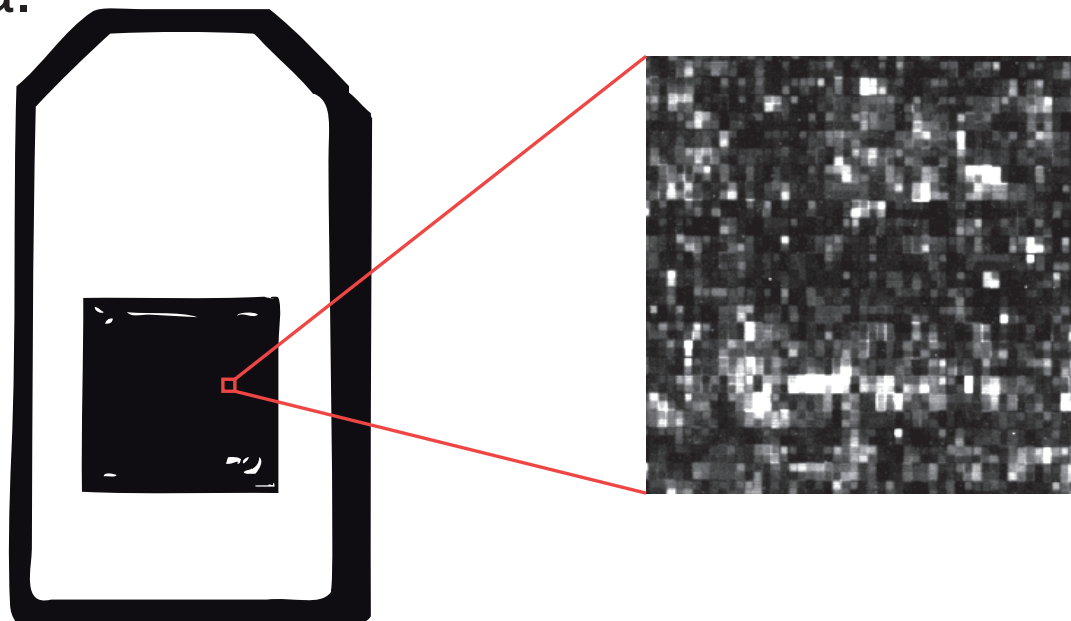
Affymetrix



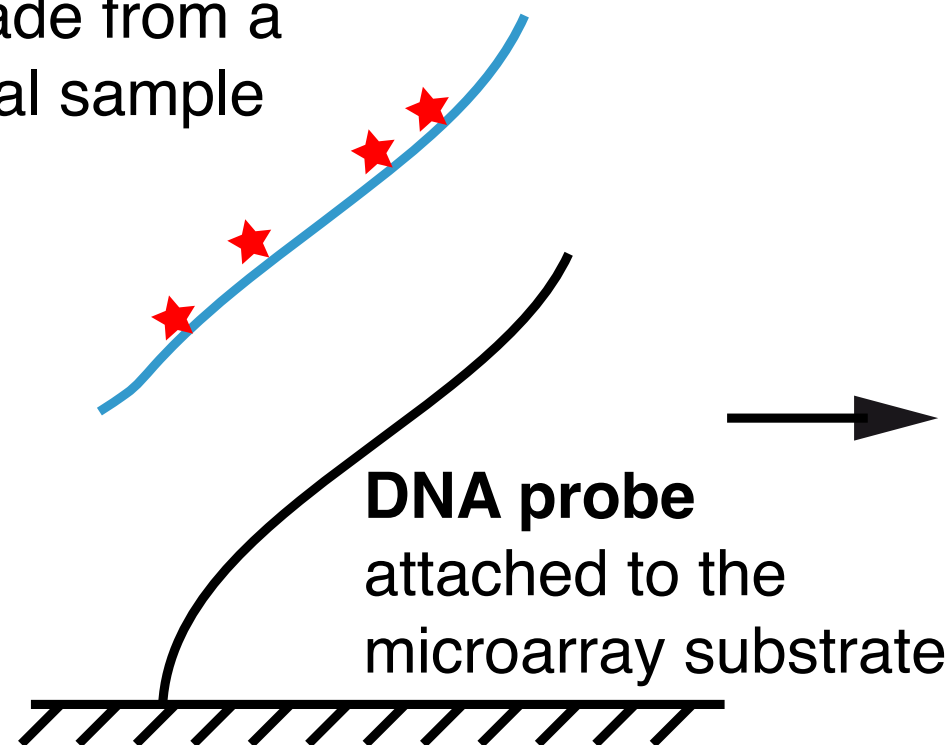
Illumina  
(24-plex)

# The gene expression microarray

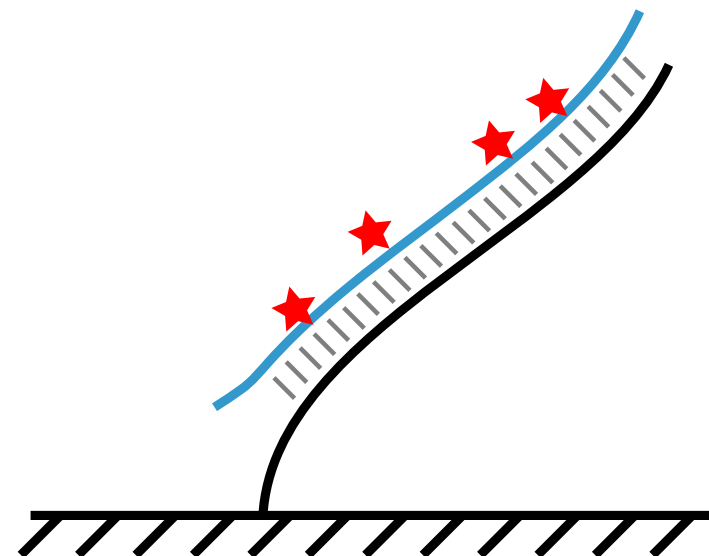
a.



**Labeled target**  
RNA made from a  
biological sample



**Specific hybridization**  
between a *complementary*  
probe and target  
detected by fluorescence



# Gene expression profile

The results from a single sample run on a microarray:

<b>Identifier</b>	<b>DDR1</b>	10.404655
	<b>RFC2</b>	6.250778
	<b>HSPA6</b>	6.221254
	<b>PAX8</b>	7.839859
	<b>GUCA1A</b>	4.019081
	<b>UBA7</b>	7.746138
	<b>THRA</b>	5.535657
	<b>PTPN21</b>	4.403469
	<b>CCL5</b>	8.516195
	<b>CYP2E1</b>	3.805621
	<b>EPHB3</b>	6.256173
	<b>ESRRA</b>	7.745937
	<b>CYP2A6</b>	5.948766
	<b>GAS6</b>	9.765667
	<b>MMP14</b>	7.721172
	<b>TRADD</b>	7.978667
	<b>FNTB</b>	4.775533
	<b>PLD1</b>	5.559748

An expression value for each “feature”



**Typically 20,000 – 50,000 genes**

# Gene expression matrix

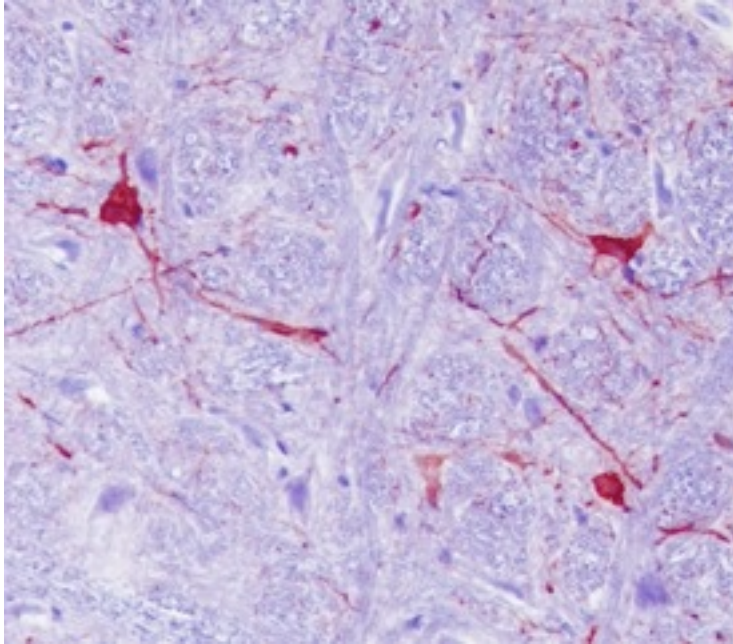
Samples →

Genes ↓

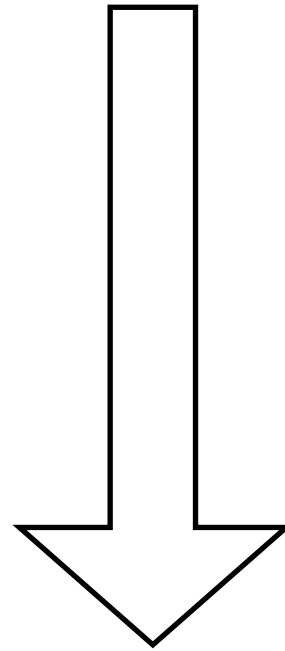
	Sample 01	Sample 02	Sample 03	Sample 04	Sample 05
<b>DDR1</b>	10.404655	10.134845	10.089483	10.317264	10.754314
<b>RFC2</b>	6.250778	6.320878	5.844462	6.108286	6.205786
<b>HSPA6</b>	6.221254	6.156638	7.332636	6.254475	6.510172
<b>PAX8</b>	7.839859	7.931699	7.846006	7.966344	7.821996
<b>GUCA1A</b>	4.019081	3.960418	4.119253	4.143439	4.022938
<b>UBA7</b>	7.746138	8.120003	7.595514	7.892662	8.025442
<b>THRA</b>	5.535657	5.262903	5.613080	5.284967	5.581082
<b>PTPN21</b>	4.403469	4.285923	5.090882	4.606929	4.657228
<b>CCL5</b>	8.516195	9.008406	5.616732	6.603475	5.941453
<b>CYP2E1</b>	3.805621	3.830757	3.923257	4.057170	3.917387
<b>EPHB3</b>	6.256173	6.167152	6.214367	6.529342	6.871130
<b>ESRRA</b>	7.745937	7.778842	7.836441	7.678031	7.753971
<b>CYP2A6</b>	5.948766	8.556330	7.456848	6.466415	9.677679
<b>GAS6</b>	9.765667	10.126754	9.601126	10.618657	10.911502
<b>MMP14</b>	7.721172	7.846023	7.954670	8.010790	8.260653
<b>TRADD</b>	7.978667	7.985578	7.581977	7.648343	8.106099
<b>FNTB</b>	4.775533	4.913803	5.138638	5.264397	5.089515
<b>PLD1</b>	5.559748	11.580453	5.525487	5.579042	5.594873

which values can be compared here?

# The process



Biological specimen

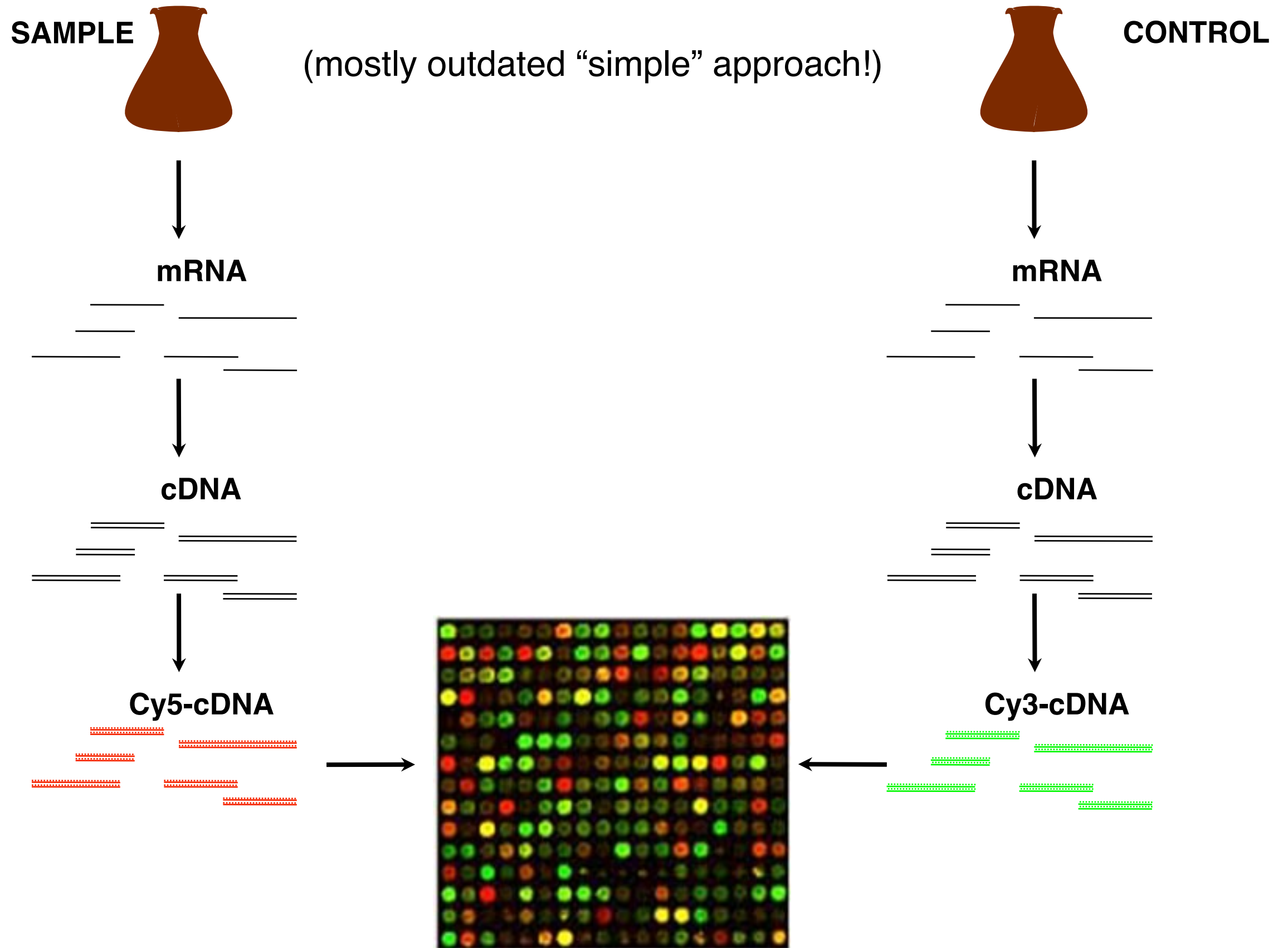


Expression profile

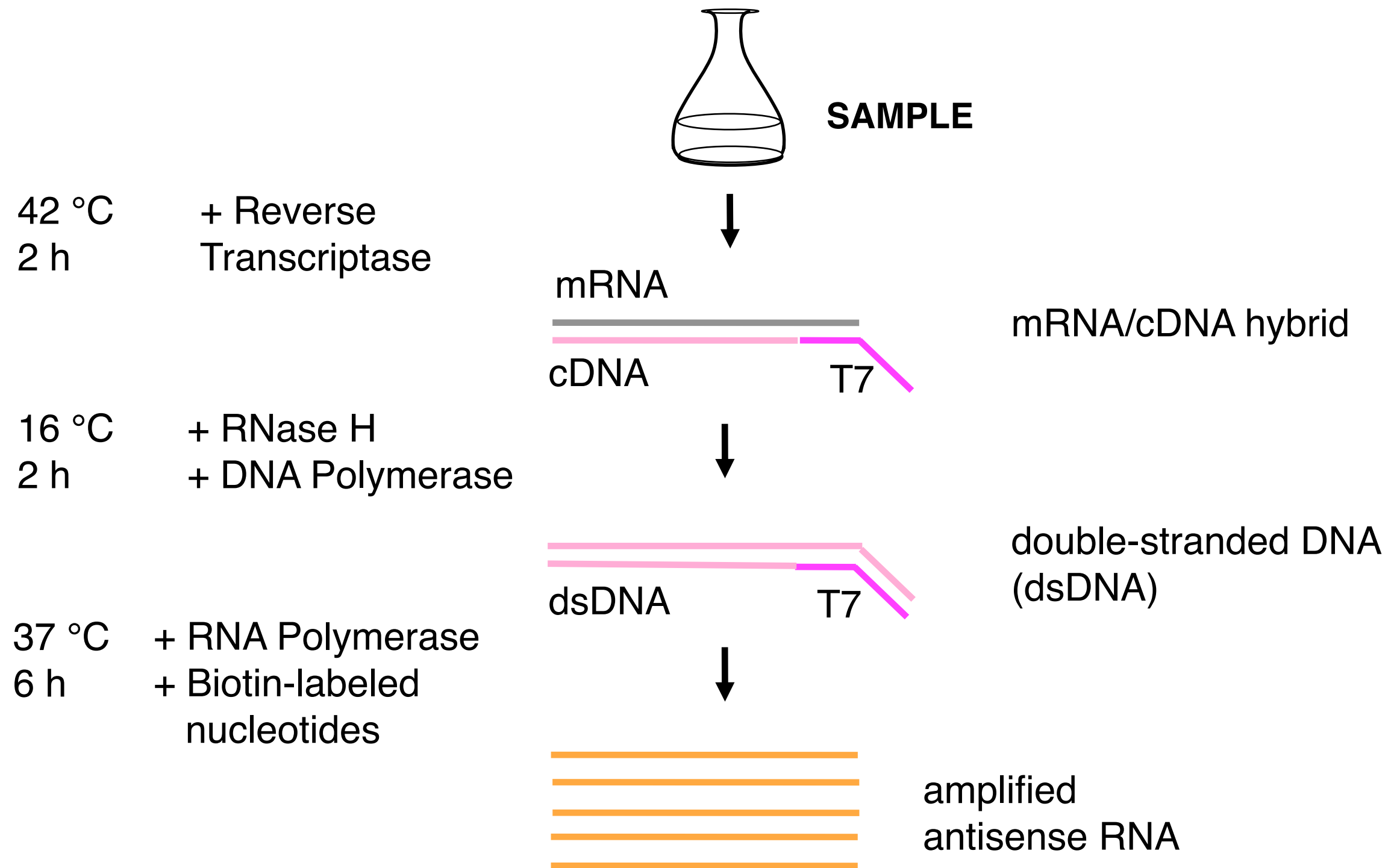
# The process

- 1. Isolate RNA from a biological specimen**
- 2. Create “labeled target” from the RNA**
- 3. Let the labeled target hybridize (incubate) on the microarray**
- 4. Wash off the unbound target**
- 5. Scan the microarray**
- 6. Analyze the data**

# Two-color labeled cDNA

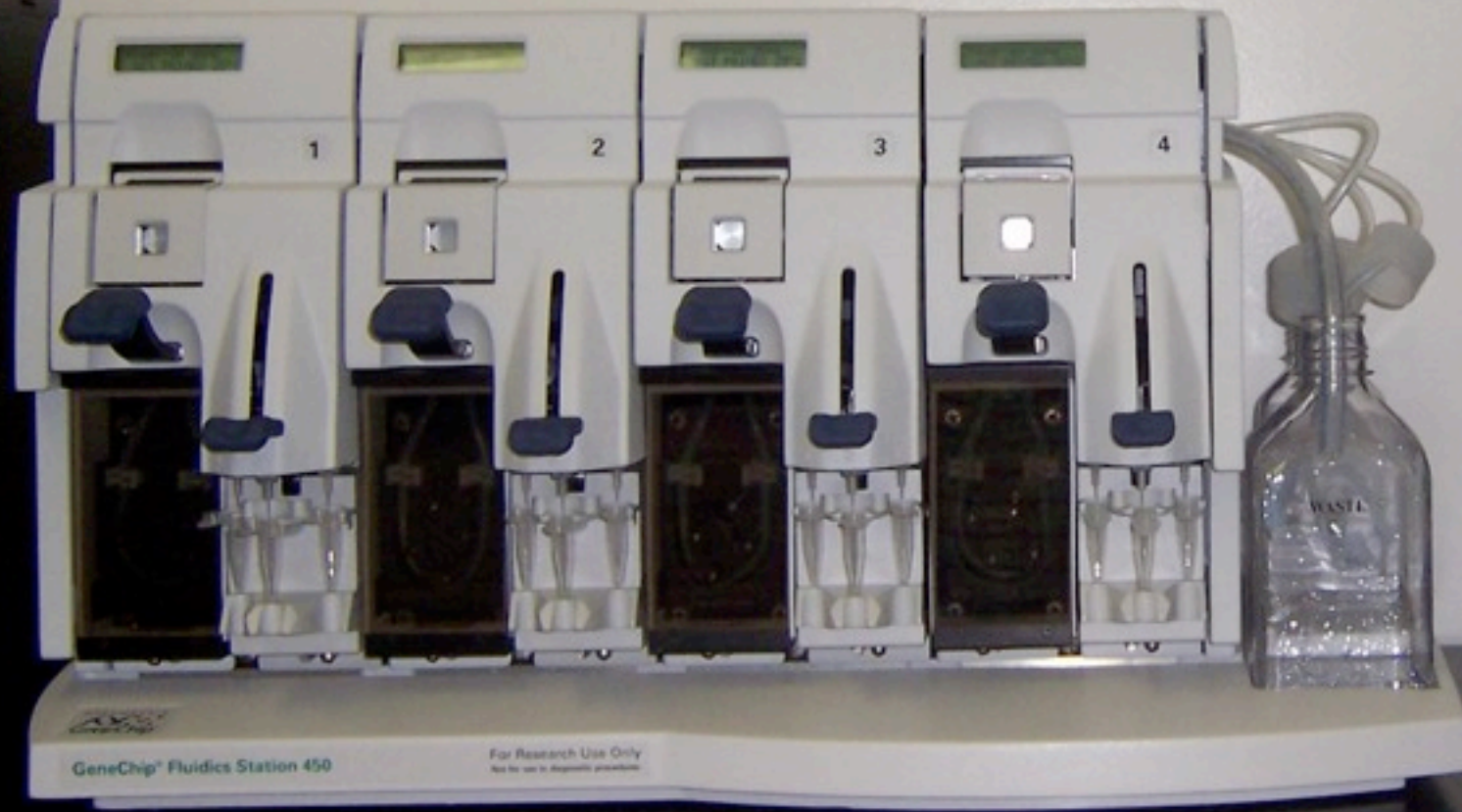


# The Eberwine protocol: for “one-color / single-channel” microarrays





# Affymetrix fluidics station and scanner

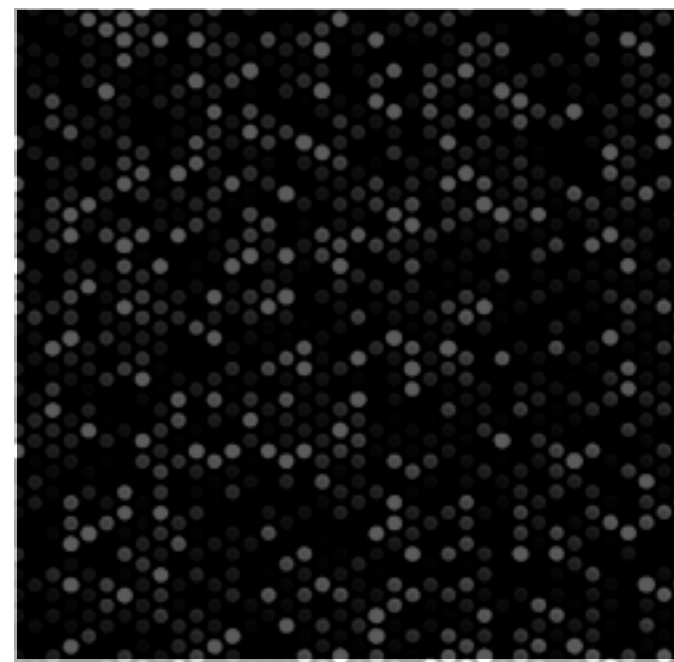
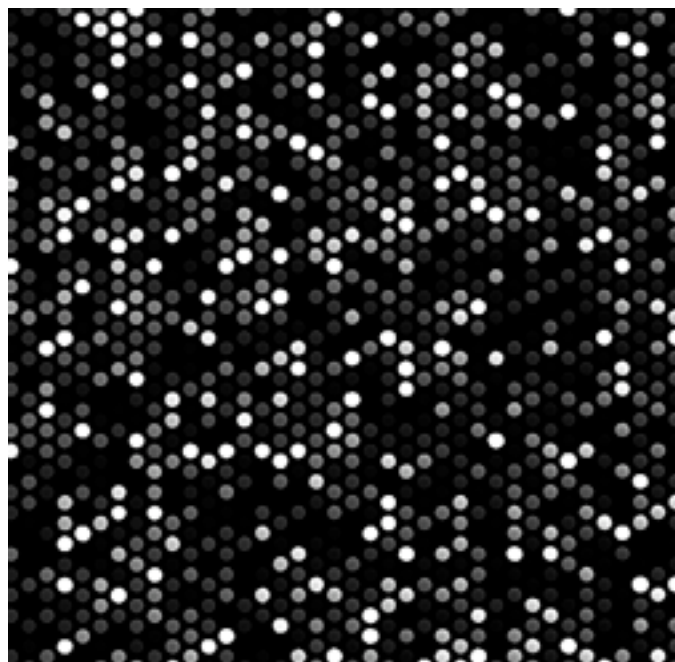


Hybridize, incubate, rinse

Scan

# Image processing and normalization

Normalization adjusts data to correct for systematic bias in intensity levels (e.g. due to different amounts of material)



# Normalized expression matrix

Samples →

Genes ↓

	Sample 01	Sample 02	Sample 03	Sample 04	Sample 05
<b>DDR1</b>	10.404655	10.134845	10.089483	10.317264	10.754314
<b>RFC2</b>	6.250778	6.320878	5.844462	6.108286	6.205786
<b>HSPA6</b>	6.221254	6.156638	7.332636	6.254475	6.510172
<b>PAX8</b>	7.839859	7.931699	7.846006	7.966344	7.821996
<b>GUCA1A</b>	4.019081	3.960418	4.119253	4.143439	4.022938
<b>UBA7</b>	7.746138	8.120003	7.595514	7.892662	8.025442
<b>THRA</b>	5.535657	5.262903	5.613080	5.284967	5.581082
<b>PTPN21</b>	4.403469	4.285923	5.090882	4.606929	4.657228
<b>CCL5</b>	8.516195	9.008406	5.616732	6.603475	5.941453
<b>CYP2E1</b>	3.805621	3.830757	3.923257	4.057170	3.917387
<b>EPHB3</b>	6.256173	6.167152	6.214367	6.529342	6.871130
<b>ESRRA</b>	7.745937	7.778842	7.836441	7.678031	7.753971
<b>CYP2A6</b>	5.948766	8.556330	7.456848	6.466415	9.677679
<b>GAS6</b>	9.765667	10.126754	9.601126	10.618657	10.911502
<b>MMP14</b>	7.721172	7.846023	7.954670	8.010790	8.260653
<b>TRADD</b>	7.978667	7.985578	7.581977	7.648343	8.106099
<b>FNTB</b>	4.775533	4.913803	5.138638	5.264397	5.089515
<b>PLD1</b>	5.559748	5.580453	5.525487	5.579042	5.594873

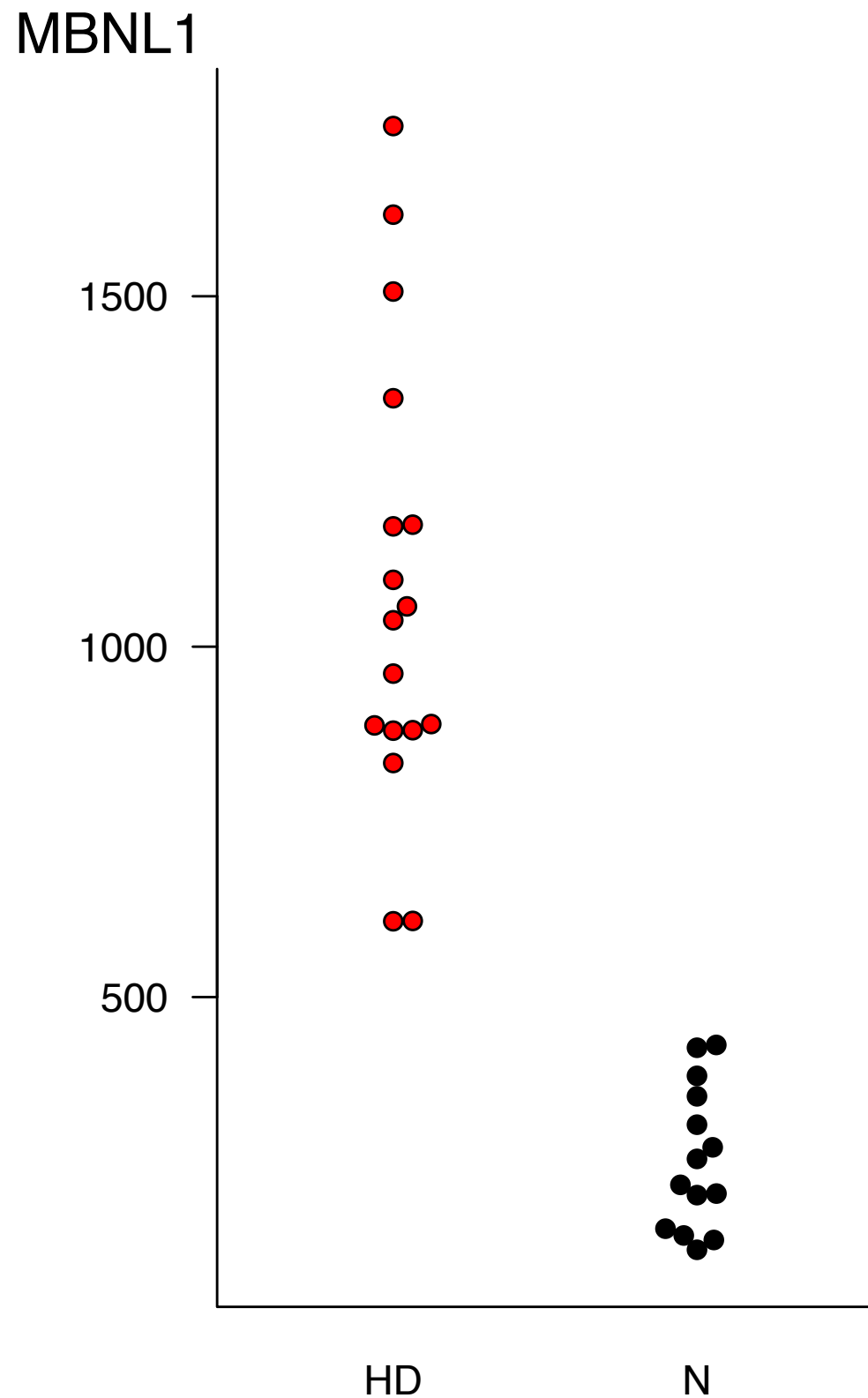
now what?

# What is the question?

Typically, we use microarrays to compare between groups:

- Disease vs. normal
- Drug treatment vs. control
- Disease A vs. Disease B
- Good prognosis vs. bad prognosis

# Comparing a gene between two groups

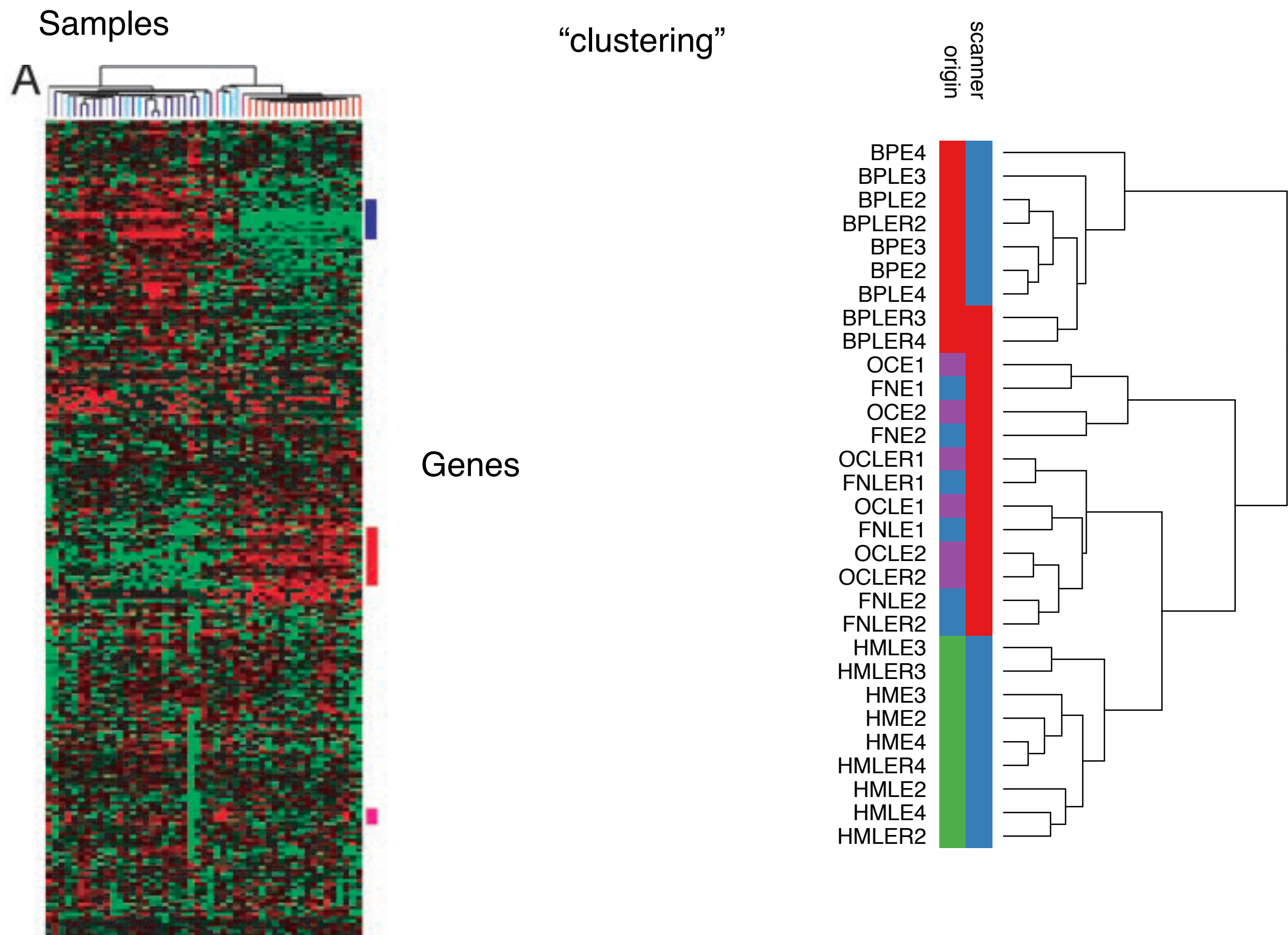


- fold change
- $\log_2$  ratio
- $P$  value (Student's  $t$  test)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$



# Visualization of microarray data



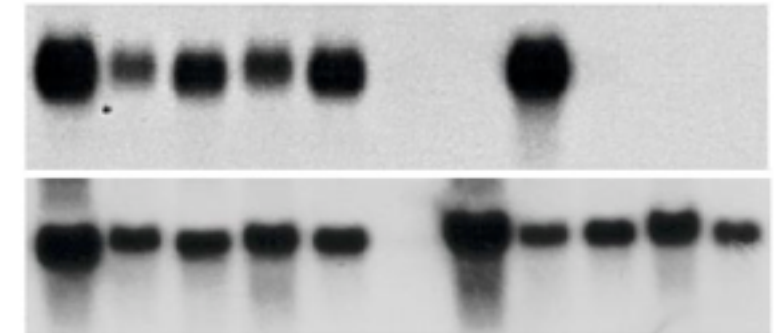
# Microarray data: what do the numbers mean

1. **No calibration:** Expression values cannot be readily translated into concentration
2. **Response curves are not parallel:** Comparison between one gene and another gene is almost meaningless
3. **Noisy at low levels** (on the log scale): Fold changes can be misleading.
4. **Saturation at high levels:** Cannot detect changes

# Evolution of gene expression (mRNA) measurements

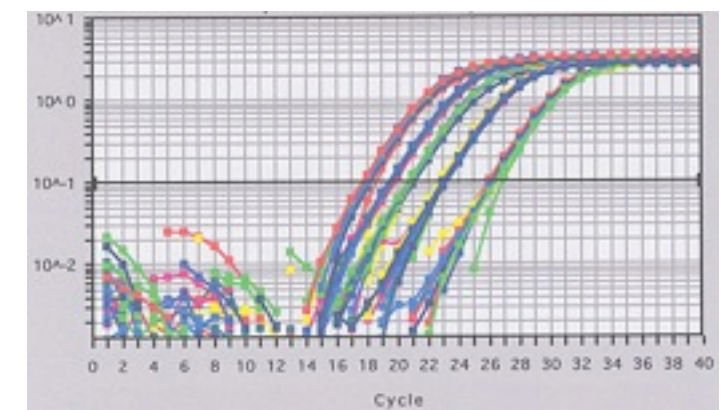
## 1. 1970s - Northern blot

- one gene at a time, not very accurate



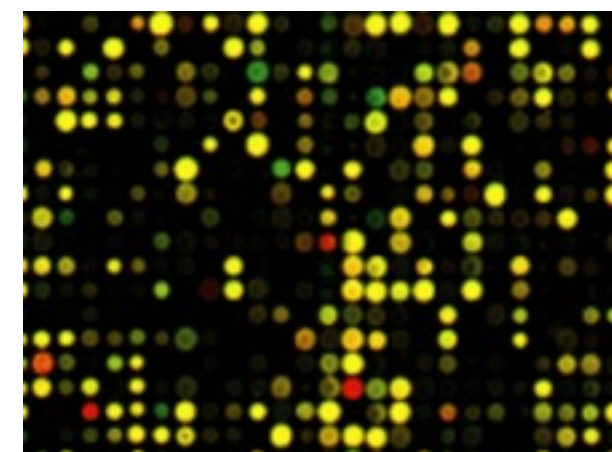
## 2. 1980s - Quantitative reverse transcriptase polymerase chain reaction (RT-PCR)

- one gene at a time, very accurate



## 3. mid-1990s - DNA microarray

- many genes at a time, medium accuracy



## 4. Late 2000s - RNA-seq

- *all* genes, higher accuracy (?)





# Other applications/types of microarrays

## DNA microarray

- **Gene expression**
- Genotyping (SNP profile)
- DNA copy number profiling (CGH, copy number)
- ChIP-Chip

## Protein microarrays

## Tissue microarrays

## Cellular microarrays

## Compound microarrays

# About the exercises

# The data set in the exercise

---

## Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease.

Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P, Bouzou B, Jensen RV, Krainc D.

*Proc Natl Acad Sci U S A.* 2005 Aug 2;102(31):11023-8.

31 samples:

- 14 normal
- 5 presymptomatic
- 12 symptomatic

Affymetrix HG-U133A arrays

Huntingon's disease:

- neurological disorder
- polyglutamine expansion in the huntingtin gene

Why search for marker of disease  
*progression* (not diagnosis)?

- assess treatment efficacy
- surrogate endpoint in drug trials

# Practicalities



Lunch break now

Exercises at 13:00